SMILES-Based Optimal Descriptors: QSAR Analysis of Fullerene-Based HIV-1 PR Inhibitors by Means of Balance of Correlations

ANDREY A. TOROPOV,[1,2] ALLA P. TOROPOVA,[1,2] EMILIO BENFENATI,[2] DANUTA LESZCZYNSKA,[3] JERZY LESZCZYNSKI[3]

[1]*Institute of Geology and Geophysics, Laboratory of Physicochemical Methods of Analysis, Khodzhibaev St. 49, 100041 Tashkent, Uzbekistan*

[2]*Istituto di Ricerche Farmacologiche Mario Negri, Laboratory of Environmental Chemistry and Toxicology, 20156, Via La Masa 19, Milano, Italy*

[3]*Department of Chemistry, Nanotoxicity Center, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson MS 39217*

Received 19 March 2009; Revised 16 April 2009; Accepted 20 April 2009
DOI 10.1002/jcc.21333
Published online 28 May 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Quantitative structure-activity relationships (QSAR) for prediction of binding affinities (pEC50, i.e., minus decimal logarithm of the 50% effective concentration) of 20 fullerene derivatives inhibitors of the HIV-1 PR (human immunodeficiency virus type 1 protease) have been developed by application of the optimal descriptors approach calculated with SMILES (simplified molecular input line entry system). The applied models were constructed by the balance of correlations. Three various splits of the experimental data into subtraining set, calibration set, and test set were examined. Comparison of classic scheme (training-test system) and the balance of correlations (subtraining-calibration-test system) show that the balance of correlations gives more robust predictions than the classic scheme for the pEC50 of the fullerene derivatives.

 J Comput Chem 31: 381–392, 2010

Key words: Fullrene; QSAR; HIV-1 PR; SMILES; optimal descriptor; correlation balance

Introduction

An application of various nanomaterials in diverse areas of industry has been increasing appreciably in the last decade. It is parallel to the investigation of fundamental properties of nanostructures by chemists, biochemists, and medicinal chemists. Usually, such experimental investigations can be extended and generalized by the results of computational studies. Among various computational approaches the quantitative structure-property/structure-activity relationships (QSPR/QSAR)-based methods by combination of experimental data with theoretical descriptors[1–12] provide useful tools to supply necessary information assisting development of novel nanomaterials[13–17] and safe applications of innovative nanotechnologies.

The molecular graphs symbolize traditional elucidation of the molecular structure in the QSPR/QSAR analysis. In many cases, the molecular graph is represented for the QSPR/QSAR analyses by the adjacency matrix. However, the molecular graph of a nanomaterial could require extremely large adjacency matrix. Simplified molecular input line entry system (SMILES)[18–20] provides a convenient alternative to molecular graph in the QSPR/QSAR analyses.[21,22]

The main problem of the QSPR/QSAR analysis is evaluation and control of the predictive ability of the developed model. Well-known procedures of the QSPR/QSAR models validation (leave-one-out, leave-many-out) deal with serious criticism.[23–26] Suggested alternatives[25,26] are based on geometric analysis of dots in coordinates of observed versus calculated (predicted) values of endpoints. In the present study we propose to use the statistical approach[13,17] as a tool to evaluate the applied model. We have examined and compared three different splits into training set and test set (traditional classic approach) as well as three splits into subtraining set, calibration set, and test set (balance of correlations[27]).

Additional Supporting Information may be found in the online version of this article.

Correspondence to: A. A. Toropov; e-mail: aatoropov@yahoo.com

Contract/grant sponsor: Marie Curie Fellowship (CHEMPREDICT); contract/grant number: 39036

Contract/grant sponsor: NSF-CREST; contract/grant number: HRD-0833178

In the last few years, biomedical applications of fullerene and its derivatives have been extensively investigated.[1,2] Durdagi et al.[1] have built QSAR for binding affinities of fullerene derivatives with the HIV-1 PR (human immunodeficiency virus type 1 aspartic protease). The aim of this study is the estimation of ability of SMILES-based optimal descriptors for QSAR analysis of the above-mentioned fullerene-based HIV-1 PR inhibitors.

Data

The binding affinity values of fullerene derivatives pEC50 were taken from work by Durdagi et al.[1] The molecular structures of these fullerene derivatives are presented in Supporting Information section. SMILES notations for these structures have been generated by ChemSketch software.[28]

Method

Optimal descriptors which are used in this study are calculated as follows:

$$\mathrm{DCW(Threshold)} = \Sigma W(S_k) \quad (1)$$

where S_k is a symbol in the SMILES notations; $W(S_k)$ represents correlation weight of the S_k. Threshold is criterion for definition (calculation) of rare symbol. If Threshold is equal to 3, then symbols (S_k) that occur in the training (or subtraining) set less than three times are classified as rare ones. Correlation weights for rare S_k is defined as $W(S_k) = 0$. Thus, the rare symbols of the SMILES have no influence on the model.

Correlation weights of active (i.e., not rare) S_k are calculated by the Monte–Carlo method optimization.[17] Two systems of the optimization have been examined in the present work.

The first is the traditional classic scheme: all available data were split into training set and test set. Correlation weights which produce as large as possible correlation coefficient between the DCW(Threshold) and pEC50 are calculated by this optimization.[10,13,17]

The second scheme is named balance of correlations: all available data were split into subtraining set, calibration set, and test set. Target function[27] of the optimization for this scheme is calculated as

$$\mathrm{CB} = R_s + R_c - \mathrm{ABS}(R_s - R_c) \times 0.1 \quad (2)$$

where R_s and R_c are correlation coefficient between the DCW(Threshold) and pEC50 for the subtraining set and calibration set, respectively. The calibration set plays role of the preliminary test set. Three splits into the subtraining set, calibration set, and test set have been examined. In case of the classic scheme, the training set contains both the subtraining set and the calibration set. Table 1 shows these three splits.

Table 1. Three Splits of the Data into the Subtraining Set, Calibration Set, and Test Set.

	Split A	Split B	Split C
Subtraining set			
	2	3	2
	3	4	4
	9	5	8
	10	7	10
	13	15	12
	15	16	16
	20	18	19
		20	20
Calibration set			
	1	1	3
	6	2	5
	7	8	6
	8	9	7
	12	10	14
	17	11	15
	18	13	18
	19		
Test set			
	4	6	1
	5	12	9
	11	14	11
	14	17	13
	16	19	17

Results and Discussion

Statistical characteristics of the models for the three splits obtained by the balance of correlations and by the classic scheme are collected in Table 2. One can see from Table 2, that balance of correlations give more accurate models for three splits. Graphically, these results obtained with the threshold ranging from 1 to 7 are shown in Figure 1.

Figure 2 shows the best model for pLD50 as well as the outlier (ID 18) that takes place in the calibration set (Split C, Probe 1, Threshold = 1). This model is characterized below:

$$\begin{aligned}
&\mathrm{pEC50} = -31.607 + 0.1247 \times \mathrm{DCW}(1)\\
&n = 8,\ r^2 = 0.9058,\ q^2 = 0.8456,\ s = 0.352, F = 58 (\text{subtraining set})\\
&n = 7, r^2 = 0.5232, R_m^2 = 0.1321, s = 1.27,\ F = 5 (\text{calibration set})\\
&n = 5, r^2 = 0.9919, R_m^2 = 0.9619, s = 0.175, F = 367 (\text{test set})
\end{aligned} \quad (3)$$

where q^2 is leave-one-out correlation coefficient value, and R_m^2 is the criterion suggested by Roy P. P. and Roy K.,[26] $R_m^2 = r^2 \times (1 - \sqrt{r^2 - r_0^2})$. The parameters r^2 and r_0^2 are squared correlation coefficient values between observed and predicted values of the test set compounds with and without the intercept respectively. For a model with good external predictability, R_m^2 value should be greater than 0.5. In the case of all 20 compounds $R_m^2 = 0.41$, but in the case of 19 compounds without the outlier (Fig. 2) the $R_m^2 = 0.559$. Thus, one can see that the model calculated with eq. (3) has external predictability according to the R_m^2 criterion, at least if the outlier is removed. It is to be noted that in spite of the presence of the outlier in the

Table 2. Statistical Characteristics of QSAR Models of the pEC_{50} for Split A, Split B, and Split C for the Threshold Ranging from 1 to 7.

			Subtraining set				Calibration set				Test set			
Threshold	Nact[a]	Probe	*n*	r^2	*s*	*F*	*n*	r^2	*s*	*F*	*n*	r^2	*s*	*F*
Split A Balance of correlations														
1	18	1	7	0.8413	0.309	27	8	0.7560	1.002	19	5	0.8379	0.642	16
		2	7	0.8177	0.332	22	8	0.8204	0.955	27	5	0.9053	0.455	29
		3	7	0.8255	0.324	24	8	0.7917	0.990	23	5	0.8964	0.599	26
Average				0.8282	0.322	24		0.7893	0.982	23		0.8799	0.565	23
2	17	1	7	0.8188	0.331	23	8	0.8160	0.989	27	5	0.9489	0.533	56
		2	7	0.8240	0.326	23	8	0.7754	1.018	21	5	0.9107	0.672	31
		3	7	0.8252	0.325	24	8	0.8214	0.964	28	5	0.9172	0.453	33
Average				0.8227	0.327	23		0.8043	0.990	25		0.9256	0.553	40
3	17	1	7	0.8234	0.326	23	8	0.8102	0.993	26	5	0.9339	0.547	42
		2	7	0.8353	0.315	25	8	0.7807	0.999	21	5	0.9402	0.617	47
		3	7	0.8078	0.341	21	8	0.8116	0.978	26	5	0.9124	0.566	31
Average				0.8222	0.327	23		0.8008	0.990	24		0.9288	0.577	40
4	16	1	7	0.8290	0.321	24	8	0.8249	0.954	28	5	0.9084	0.422	30
		2	7	0.8303	0.320	24	8	0.8252	0.959	28	5	0.9086	0.399	30
		3	7	0.8352	0.315	25	8	0.7802	0.961	21	5	0.8946	0.483	25
Average				0.8315	0.319	25		0.8101	0.958	26		0.9039	0.435	28
5	15	1	7	0.9197	0.220	57	8	0.5180	1.775	6	5	0.0035	1.769	0
		2	7	0.9165	0.225	55	8	0.5055	1.906	6	5	0.0104	1.673	0
		3	7	0.9295	0.206	66	8	0.5042	1.701	6	5	0.0222	1.848	0
Average				0.9219	0.217	59		0.5092	1.794	6		0.0120	1.763	0
6	15	1	7	0.9160	0.225	55	8	0.5412	1.846	7	5	0.0001	1.677	0
		2	7	0.9357	0.197	73	8	0.5359	1.890	7	5	0.0083	1.796	0
		3	7	0.9232	0.215	60	8	0.5413	1.796	7	5	0.0039	1.692	0
Average				0.9250	0.212	62		0.5395	1.844	7		0.0041	1.722	0
7	14	1	7	0.9531	0.168	102	8	0.5387	2.008	7	5	0.0031	1.734	0
		2	7	0.9245	0.213	61	8	0.5557	1.879	8	5	0.0240	1.354	0
		3	7	0.9003	0.245	45	8	0.5716	1.649	8	5	0.0033	1.474	0
Average				0.9260	0.209	69		0.5554	1.845	8		0.0101	1.521	0
Split A Classic scheme														
1	22	1	15	0.8412	0.432	69					5	0.6149	1.753	5
		2	15	0.8353	0.440	66					5	0.6211	1.794	5
		3	15	0.8440	0.428	70					5	0.5290	1.864	3
Average				0.8402	0.433	68						0.5883	1.804	4
2	18	1	15	0.7490	0.543	39					5	0.6835	1.336	6
		2	15	0.7525	0.539	40					5	0.6789	1.593	6
		3	15	0.7570	0.534	41					5	0.6459	1.606	5
Average				0.7529	0.539	40						0.6694	1.512	6
3	18	1	15	0.7681	0.522	43					5	0.6237	1.712	5
		2	15	0.7695	0.520	43					5	0.6145	1.788	5
		3	15	0.7556	0.536	40					5	0.6727	1.582	6
Average				0.7644	0.526	42						0.6369	1.694	5
4	18	1	15	0.7539	0.538	40					5	0.6456	1.549	5
		2	15	0.7486	0.543	39					5	0.6752	1.607	6
		3	15	0.7585	0.533	41					5	0.6560	1.535	6
Average				0.7537	0.538	40						0.6590	1.563	6
5	17	1	15	0.7572	0.534	41					5	0.6318	1.549	5
		2	15	0.7613	0.529	41					5	0.6431	1.609	5
		3	15	0.7591	0.532	41					5	0.6460	1.576	5
Average				0.7592	0.532	41						0.6403	1.578	5
6	16	1	15	0.7589	0.532	41					5	0.6519	1.557	6
		2	15	0.7568	0.534	40					5	0.6473	1.583	6
		3	15	0.7572	0.534	41					5	0.6528	1.530	6
Average				0.7576	0.533	41						0.6507	1.557	6
7	16	1	15	0.7586	0.532	41					5	0.6470	1.635	5
		2	15	0.7628	0.528	42					5	0.6414	1.672	5
		3	15	0.7613	0.529	41					5	0.6510	1.637	6
Average				0.7609	0.530	41						0.6465	1.648	5

(*Continued*)

Table 2. (*Continued*).

			Subtraining set				Calibration set				Test set			
Threshold	Nact[a]	Probe	*n*	r^2	*s*	*F*	*n*	r^2	*s*	*F*	*n*	r^2	*s*	*F*
Split B Balance of correlations														
1	19	1	8	0.7778	0.605	21	7	0.9127	0.768	52	5	0.6740	0.752	6
		2	8	0.7911	0.587	23	7	0.8977	0.716	44	5	0.7880	0.820	11
		3	8	0.7814	0.601	21	7	0.9484	0.944	92	5	0.5976	0.699	4
Average				0.7834	0.598	22		0.9196	0.809	63		0.6866	0.757	7
2	18	1	8	0.7778	0.605	21	7	0.9433	0.945	83	5	0.5894	0.696	4
		2	8	0.7872	0.592	22	7	0.8946	0.776	42	5	0.7832	0.797	11
		3	8	0.7719	0.613	20	7	0.9696	0.954	160	5	0.6997	0.656	7
Average				0.7790	0.604	21		0.9358	0.892	95		0.6907	0.716	7
3	17	1	8	0.7583	0.631	19	7	0.9677	0.826	150	5	0.9808	0.810	153
		2	8	0.7605	0.629	19	7	0.9502	0.705	95	5	0.9579	0.797	68
		3	8	0.7692	0.617	20	7	0.9568	0.805	111	5	0.9797	0.836	145
Average				0.7626	0.626	19		0.9582	0.778	119		0.9728	0.814	122
4	17	1	8	0.7766	0.607	21	7	0.9563	0.834	109	5	0.9519	0.833	59
		2	8	0.7634	0.625	19	7	0.9493	0.798	94	5	0.9790	0.799	140
		3	8	0.7560	0.634	19	7	0.9717	0.898	172	5	0.9613	0.798	75
Average				0.7653	0.622	20		0.9591	0.843	125		0.9641	0.810	91
5	15	1	8	0.6917	0.713	13	7	0.7549	1.057	15	5	0.1339	0.998	0
		2	8	0.6989	0.705	14	7	0.7667	1.023	16	5	0.1158	1.011	0
		3	8	0.7000	0.704	14	7	0.7462	1.068	15	5	0.0943	0.976	0
Average				0.6968	0.707	14		0.7559	1.050	16		0.1147	0.995	0
6	15	1	8	0.7044	0.698	14	7	0.7558	1.063	15	5	0.0430	0.967	0
		2	8	0.6895	0.716	13	7	0.7537	1.061	15	5	0.1589	0.987	1
		3	8	0.7011	0.702	14	7	0.7529	1.045	15	5	0.0865	0.983	0
Average				0.6983	0.705	14		0.7541	1.056	15		0.0961	0.979	0
7	15	1	8	0.6922	0.713	13	7	0.7534	1.057	15	5	0.1456	0.991	1
		2	8	0.6897	0.715	13	7	0.7613	1.048	16	5	0.0924	0.983	0
		3	8	0.7209	0.679	15	7	0.7484	1.070	15	5	0.0534	1.007	0
Average				0.7009	0.702	14		0.7544	1.058	15		0.0971	0.994	0
Split B Classic scheme														
1	19	1	15	0.8378	0.438	67					5	0.5163	1.320	3
		2	15	0.8380	0.438	67					5	0.4904	1.338	3
		3	15	0.8482	0.424	73					5	0.1518	1.340	1
Average				0.8414	0.433	69						0.3862	1.332	2
2	18	1	15	0.8378	0.438	67					5	0.3063	1.237	1
		2	15	0.8385	0.437	67					5	0.6079	1.300	5
		3	15	0.8398	0.436	68					5	0.4074	1.219	2
Average				0.8387	0.437	68						0.4405	1.252	3
3	18	1	15	0.8459	0.427	71					5	0.4614	1.346	3
		2	15	0.8423	0.432	69					5	0.4259	1.415	2
		3	15	0.8490	0.423	73					5	0.3203	1.349	1
Average				0.8457	0.427	71						0.4026	1.370	2
4	17	1	15	0.8430	0.431	70					5	0.2926	1.530	1
		2	15	0.8450	0.428	71					5	0.3279	1.382	1
		3	15	0.8369	0.440	67					5	0.5129	1.330	3
Average				0.8416	0.433	69						0.3778	1.414	2
5	17	1	15	0.8466	0.426	72					5	0.5257	1.391	3
		2	15	0.8448	0.429	71					5	0.4025	1.431	2
		3	15	0.8528	0.418	75					5	0.2261	1.350	1
Average				0.8481	0.424	73						0.3848	1.391	2
6	17	1	15	0.8441	0.430	70					5	0.1284	1.229	0
		2	15	0.8445	0.429	71					5	0.5423	1.415	4
		3	15	0.8403	0.435	68					5	0.4767	1.453	3
Average				0.8430	0.431	70						0.3825	1.366	2
7	17	1	15	0.8437	0.430	70					5	0.5804	1.350	4
		2	15	0.8402	0.435	68					5	0.3865	1.247	2
		3	15	0.8448	0.429	71					5	0.4961	1.538	3
Average				0.8429	0.431	70						0.4877	1.379	3

(*Continued*)

Table 2. (*Continued*).

			Subtraining set				Calibration set				Test set			
Threshold	Nact[a]	Probe	*n*	r^2	*s*	*F*	*n*	r^2	*s*	*F*	*n*	r^2	*s*	*F*
Split C Balance of correlations														
1[b]	19	1	8	0.9058	0.352	58	7	0.5232	1.275	5	5	0.9919	0.162	367
		2	8	0.8912	0.378	49	7	0.4443	1.327	4	5	0.9553	0.341	64
		3	8	0.8873	0.385	47	7	0.5319	1.337	6	5	0.9359	0.212	44
Average				0.8948	0.372	51		0.4998	1.313	5		0.9610	0.238	158
2	19	1	8	0.8603	0.429	37	7	0.4905	1.226	5	5	0.9779	0.182	133
		2	8	0.9077	0.348	59	7	0.5107	1.292	5	5	0.9649	0.220	82
		3	8	0.8213	0.485	28	7	0.4812	1.280	5	5	0.9661	0.198	85
Average				0.8631	0.420	41		0.4941	1.266	5		0.9696	0.200	100
3	18	1	8	0.9241	0.316	73	7	0.5786	1.263	7	5	0.4372	1.018	2
		2	8	0.8429	0.454	32	7	0.5062	1.201	5	5	0.9384	0.236	46
		3	8	0.9154	0.333	65	7	0.6054	1.629	8	5	0.2969	1.462	1
Average				0.8941	0.368	57		0.5634	1.364	7		0.5575	0.905	16
4	16	1	8	0.7903	0.525	23	7	0.4892	1.085	5	5	0.8041	0.358	12
		2	8	0.8784	0.400	43	7	0.5183	1.184	5	5	0.7018	0.478	7
		3	8	0.9026	0.358	56	7	0.5004	1.220	5	5	0.7932	0.411	12
Average				0.8571	0.428	41		0.5026	1.163	5		0.7664	0.416	10
5	16	1	8	0.8858	0.387	47	7	0.5039	1.140	5	5	0.7997	0.395	12
		2	8	0.8351	0.466	30	7	0.6249	1.044	8	5	0.1842	1.313	1
		3	8	0.7753	0.543	21	7	0.4712	1.103	4	5	0.7842	0.375	11
Average				0.8320	0.465	33		0.5333	1.096	6		0.5894	0.694	8
6	16	1	8	0.8879	0.384	48	7	0.4932	1.130	5	5	0.8106	0.398	13
		2	8	0.8171	0.490	27	7	0.6967	0.937	11	5	0.1141	1.593	0
		3	8	0.8786	0.400	43	7	0.5204	1.124	5	5	0.7747	0.405	10
Average				0.8612	0.425	39		0.5701	1.063	7		0.5664	0.799	8
7	15	1	8	0.9002	0.362	54	7	0.4865	1.955	5	5	0.1683	1.328	1
		2	8	0.8632	0.424	38	7	0.4877	1.626	5	5	0.2809	0.974	1
		3	8	0.8283	0.475	29	7	0.4903	1.577	5	5	0.2173	1.105	1
Average				0.8639	0.420	40		0.4881	1.719	5		0.2222	1.136	1
Split C Classic scheme														
1	22	1	15	0.8832	0.401	98					5	0.2945	1.612	1
		2	15	0.8816	0.404	97					5	0.2562	1.683	1
		3	15	0.8600	0.439	80					5	0.5280	1.211	3
Average				0.8749	0.415	92						0.3596	1.502	2
2	19	1	15	0.8650	0.431	83					5	0.1068	2.128	0
		2	15	0.8716	0.421	88					5	0.0893	2.321	0
		3	15	0.8744	0.416	90					5	0.0925	2.263	0
Average				0.8703	0.423	87						0.0962	2.237	0
3	18	1	15	0.8589	0.441	79					5	0.1191	2.108	0
		2	15	0.8399	0.470	68					5	0.1586	1.879	1
		3	15	0.8730	0.418	89					5	0.0925	2.245	0
Average				0.8572	0.443	79						0.1234	2.078	0
4	18	1	15	0.8670	0.428	85					5	0.1106	2.165	0
		2	15	0.8540	0.448	76					5	0.1232	2.072	0
		3	15	0.8702	0.423	87					5	0.1119	2.170	0
Average				0.8637	0.433	83						0.1152	2.136	0
5	17	1	15	0.8390	0.471	68					5	0.1775	1.727	1
		2	15	0.8458	0.461	71					5	0.1473	1.816	1
		3	15	0.8348	0.477	66					5	0.1698	1.684	1
Average				0.8399	0.470	68						0.1649	1.742	1
6	16	1	15	0.8357	0.476	66					5	0.1846	1.689	1
		2	15	0.8245	0.492	61					5	0.1900	1.615	1
		3	15	0.8342	0.478	65					5	0.1788	1.655	1
Average				0.8315	0.482	64						0.1845	1.653	1
7	16	1	15	0.8380	0.472	67					5	0.1677	1.713	1
		2	15	0.8317	0.481	64					5	0.2000	1.631	1
		3	15	0.8342	0.478	65					5	0.1852	1.665	1
Average				0.8346	0.477	66						0.1843	1.670	1

[a]The number of active attributes which are not rare for given threshold.
[b]The best prediction.

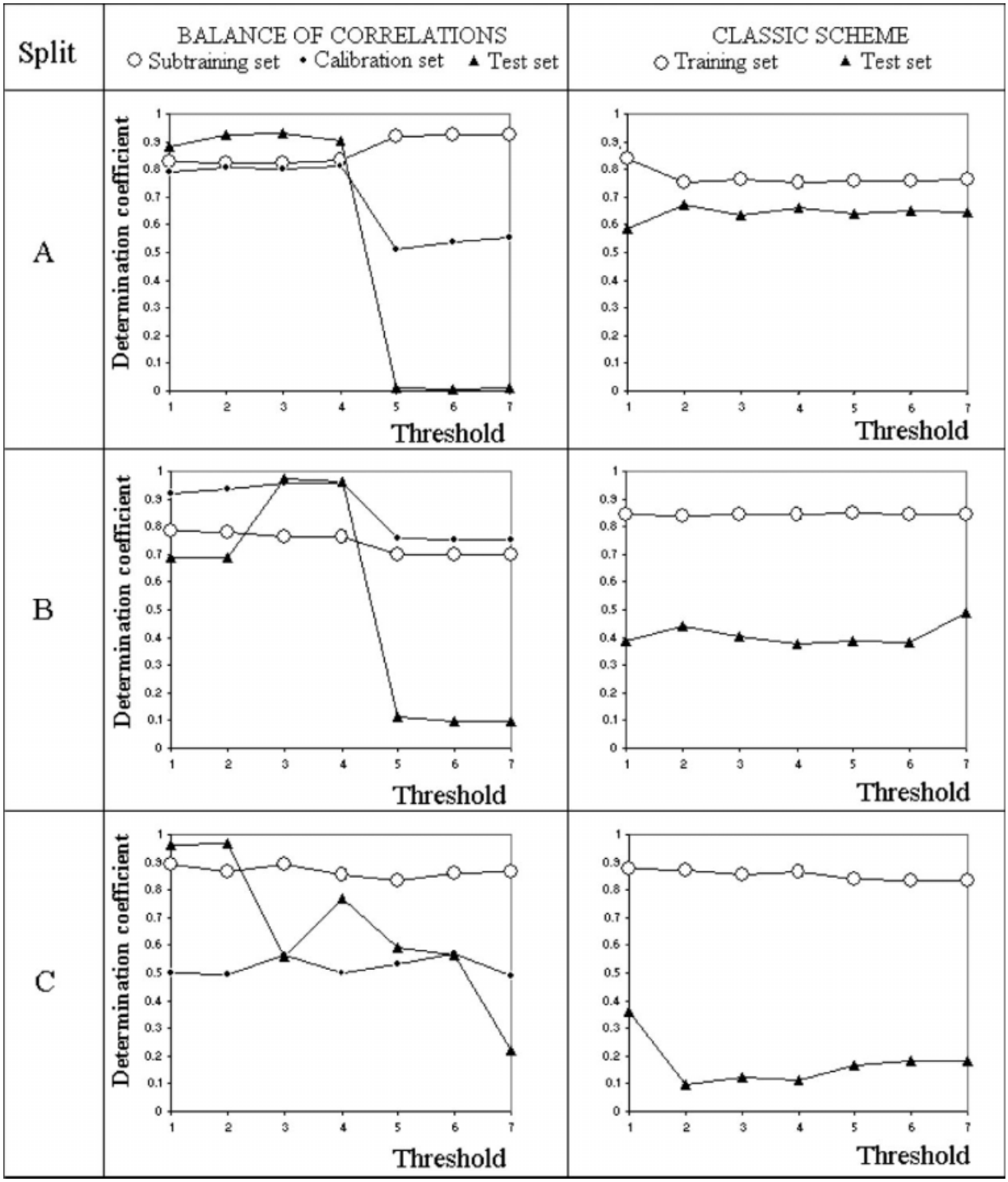

Figure 1. Comparison of the classic scheme of QSAR modeling (training-test system) and the balance of correlations (subtraining set-calibration set-test set) for three random splits into the subtraining set, calibration set, and the test set.

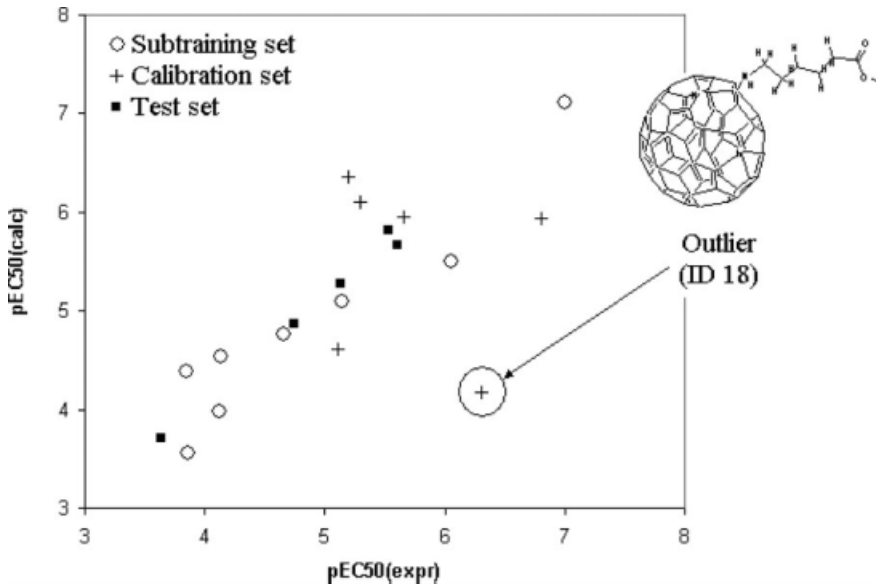

Figure 2. Graphical representation of the model calculated with eq. (3).

calibration set the prediction for the external test set is quite satisfactory. It should be noted that in the cases of Split A and Split B the compound with ID 18 (Table 2, Fig. 2) is not an outlier. Furthermore, the average value of the R_m^2 for calibration set and test set [i.e., $0.5 \times (0.1321 + 0.9619) = 0.547$] is larger than 0.5.

Table 3 contains the correlation weights for calculation with eq. (3). An example of the DCW(1) calculation is demonstrated in the Supporting Information section. Experimental and calculated values using eq. (3) of pEC50 are displayed in Table 4.

Table 5 contains the details of distribution of S_k into subtraining set, calibration set, and test set. One can see from Table 5 that number of SMILES that contain S_k in majority cases is the same, whereas the total numbers of S_k in different splits vary. Thus, the number of S_k in these splits defines the statistical quality of these models.

In case of the split C, the statistical quality of the model [eq. (3)] for the calibration set is poor ($r^2 \approx 0.5$). Nevertheless, the prediction without the calibration set (solely with the subtraining set, $n = 8$) is not better: in this case prediction for the external test set is characterized by $r^2 < 0.5$. In other words, the use of the calibration set improves statistical characteristics of the prediction, despite poor statistical characteristics for the calibration set.

Statistical quality of models for pEC50 described by Durdagi et al.[1] are excellent ($r^2 = 0.997$). However, in fact, the only cross-validated r_{cv}^2 values are presented in the report ($r_{cv}^2 \approx 0.55$). According to ref. 1, the last three compound were used as a test set (18, 19, and 20), consequently, these models are characterized (for the test set) by $n = 3$, $r^2 = 0.8347$ in case of the comparative molecular field analysis (CoMFA) and $n = 3$, $r^2 = 0.9699$ in case of comparative molecular similarity indices analysis (CoMSIA). The comparison of these results with the SMILES-based models (Table 2) indicates that SMILES-based models could be considered as a valuable alternative.

It should be noted that models described in the ref. 1 are based on the 3D-descriptors (three-dimensional), whereas, models examined here are based on 2D-descriptors (two-dimensional), hence using the described approach one can carry out the QSAR analysis using the Internet available databases[29,30] which contain SMILES notations.

Table 3. Correlation Weights for Calculation DCW(1) Used in eq. (3).

S_k	$W(S_k)$
%	1.2537874
(	0.4971701
/	−1.0048547
0	4.2539384
1	2.8730463
2	3.5045661
3	−2.1240763
4	3.7482791
5	2.0042684
6	3.0046478
7	1.9980735
8	3.3096573
9	2.4389246
=	1.8123507
@	0.0
C	−1.6270157
H	0.0
N	−2.5013376
O	−0.8140661
[	0.0
\	−2.5038693
c	−0.1291104

Split C, Probe 1 of the Optimization, Threshold = 1.

Table 4. Experimental and Calculated with eq. (3) Values of the Activity (pEC50) for 20 Fullerene-Based HIV-1 Inhibitors.

ID	SMILES	DCW(1)	pEC50	
			Expr	Calc
	Subtraining set			
2	CN%22CCC%27%24c1c7c%33C=6C4=C%31C=3c%33c2c1c%26c% 29c%18c2C=3C=%17C=%32C=%16C%12=C%15C(=C4C5=C%14C=%30C8=C5C=6C7=C%23C8=C%21C=9C=% 30C%13=C%11C=9C=%25C%20C=%10C%19=C(C(C=%10%11)=C%12C%13=C%14%15)C=%16C=%17C%18=C%19C% 29%28CCN(C)CC%20%28C(C=%25C%21C%23%24C%22)=C%26%27)C%31=%32	289.8103391	4.140	4.532
4	OC(c1ccccc1)C%11(c2ccccc2)C%21%12C5=C%28C%20=C%19C%29=C%18C%31=C%17C=%32C=4C=3C=%16C%13=C7C=6C=3C=%22C=4C%34=C%23C=%25C=%33C%30=C(C%26=C5C=%24C%10C=9C%27=C(C=6C8=C7C=%14C(=C(C8=9)C%10%11%12)C(C%15C%19=C%18C(=C%13C=%14%15)C=%16%17)=C%20%21)C=%22C%23=C%27C=%24C=%25%26)C%28=C%29C%30=C%31C=%32C=%33%34	310.4704712	7.000	7.109
8	O=C(O)C(N)CCCCNC%31%30C%13=C1C%15=C%18C%26C1=C%28C%31C%10=C%29C=%20C9=C%21C2=C6C%22=C%24C=5C%17=C4C%16=C%14C8=C%12C=3C=%11C(=C2C=7C=3C8=C4C=5C6=7)C9=C%10C=%11C%30C%12=C%13C%14=C%15C%16=C%19C%17=C%23C(=C%18%19)C=%27C%25=C(C=%20C%21=C%22C%25=C%23%24)C(C%26=%27)=C%28%29	285.4211043	4.120	3.985
10	O=C(O)CC(CC(=O)O)NC(=O)NC%31%30C%13=C1C%15=C%18C%26C1=C%28C%31C%10=C%29C=%20C9=C%21C2=C6C%22=C%24C=5C%17=C4C%16=C%14C8=C%12C=3C=%11C(=C2C=7C=3C8=C4C=5C6=7)C9=C%10C=%11C%30C%12=C%13C%14=C%15C%16=C%19C%17=C%23C(=C%18%19)C=%27C%25=C(C=%20C%21=C%22C%25=C%23%24)C(C%26=%27)=C%28%29	288.5922878	3.850	4.380
12	O=C(O)CO\N=C%22\CC%12%23C9C=4C=3C%21=C%26C%20=C%19C%27=C%18C%29=C%17C=%30C=2C=1C=%16C%13=C7C=6C=1C=%24C=2C%32=C5C=%25C=%31C%28=C(C=3C=%25C=4C%11=C5C=%24C=%10C=6C8=C7C=%14C(=C%12C8=C9C=%10%11)C(C%15C%19=C%18C(=C%13C=%14%15)C=%16%17)=C%20C%21%23C(OC)C%22)C%26=C%27C%28=C%29C=%30C=%31%32	297.6145647	6.050	5.506
16	O=C(O)CO\N=C%29/C=CC%23%31C%21=C3C2=C%30C=%28C=%32C%27=C1C8=C7C=6C5=C1C=%32C2=C4C3C%20=C%19C(=C45)C=6C=%11C=%12C7=C%14C8=C%26C%15=C%25C=%24C9=C%16C=%13C%17=C%10C9=C(C%23=C%22C%10=C%18C(=C%19C=%11C(C=%12C=%13C%14=C%15%16)=C%17%18)C%20=C%21%22)C=%24C(C=%28C%25=C%26%27)C%30%31C%29	294.3052583	5.140	5.093
19	O=C(O)CO\N=C1/CCCC%28(C)C1C%17%27C9C=%11C=%13C%26=C%24C=%29C%14=C%30C=2C=%12C%15=C5C=2C%32=C%22C=4C=3C=%21C%18=C7C=6C=3C(C=45)=C%16C=%10C=6C8=C7C=%19C(=C%17C8=C9C=%10C(C=%11C=%12C=%13%14)=C%15%16)C=%25C%20C%23=C%33C(=C%18C=%19%20)C=%21C%22=C%31C%33=C(C%23=C%24C=%25C%26%27%28)C=%29C%30=C%31%32	291.6792238	4.660	4.765

(*Continued*)

Table 4. (*Continued*).

ID	SMILES	DCW(1)	pEC50 Expr	pEC50 Calc
20	C1COCCOCCOCCOCCOc2cc(ccc2O1)C%23(c3ccccc3)C%34%22C%12=C%32C=%25C%11=C%26C4=C8C%27=C%29C=7C%17=C6C%16=C%15C%10=C%14C=5C=%13C(=C4C=9C=5C%10=C6C=7C8=9)C%11=C%12C=%13C%21C%14=C%20C%15=C%19C%16=C%18C%17=C%28C%24=C%18C%35=C%19C(=C%20C%21%22%23)C=%33C%35C%31=C%24C%30=C(C=%25C%26=C%27C%30=C%28%29)C%31=C%32C=%33%34	281.9801843	3.860	3.556
	Calibration set			
3	O=C%27CC%33%28C3=C5c1c%26c%23c%30c2c1C=4C6=C7C2=C%31C8=C%17C%29=C%22C=%16C%21=C%15C%20=C%13C=%12C=%19C(=C3C%11=C(C=45)C=%10C6=C9C7=C8C=%18C%14=C9C(C=%10C%11=%12)=C%13C%14=C%15C=%16C%17=%18)C%33C%24=C(C=%19%20)C%21C%25%32CC(=O)CC(OC)C%22%32c(c%23C(=C%24%25)C%26%28C(OC)C%27)c%29c%30%31	290.4677286	5.110	4.614
5	NCCc1ccc(cc1)C%11(c2ccc(CCN)cc2)C%21%12C5=C%28C%20=C%19C%29=C%18C%31=C%17C=%32C=4C=3C=%16C%13=C7C=6C=3C=%22C=4C%34=C%23C=%25C=%33C%30=C(C%26=C5C=%24C%10C=9C%27=C(C=6C8=C7C=%14C(=C(C8=9)C%10%11%12)C(C%15C%19=C%18C(=C%13C=%14%15)C=%16%17)=C%20%21)C=%22C%23=C%27C=%24C=%25%26)C%28=C%29C%30=C%31C=%32C=%33%34	302.3951552	5.300	6.102
6	CC(C)[C@H]%26CC(O)[C@H](C(C)C)C%17%30C%14C=1C%27=C2C=%32C=1C%16=C%31C=5C=%32C=6C9=C2C%28=C%10C%24=C%23C8=C%22C=7C=4C=3C=%21C%18=C%12C=%11C=3C(C=4C=5C=6C=7C8=C9%10)=C%31C=%15C=%11C%13=C%12C=%19C(=C%17C%13=C%14C=%15%16)C=%25C%20C(=C%23C(=C%18C=%19%20)C=%21%22)C%24=C%29C=%25C%26%30C%27=C%28%29?	301.0487711	6.800	5.934
7	O=C(O)CCC(CCCC\C=C\CCCCCCCC)NC%31%30C%13=C1C%15=C%18C%26C1=C%28C%31C%10=C%29C=%20C9=C%21C2=C6C%22=C%24C=5C%17=C4C%16=C%14C8=C%12C=3C=%11C(=C2C=7C=3C8=C4C=5C6=7)C9=C%10C=%11C%30C%12=C%13C%14=C%15C%16=C%19C%17=C%23C(=C%18%19)C=%27C%25=C(C=%20C%21=C%22C%25=C%23%24)C(C%26=%27)=C%28%29	265.2028656	2.890	1.464
14	O=C(O)CCC(=O)OCC%26CCC(C)C%17%30C%14C=1C%27=C2C=%32C=1C%16=C%31C=5C=%32C=6C9=C2C%28=C%10C%24=C%23C8=C%22C=7C=4C=3C=%21C%18=C%12C=%11C=3C(C=4C=5C=6C=7C8=C9%10)=C%31C=%15C=%11C%13=C%12C=%19C(=C%17C%13=C%14C=%15%16)C=%25C%20C(=C%23C(=C%18C=%19%20)C=%21%22)C%24=C%29C=%25C%26%30C%27=C%28%29	301.2369340	5.660	5.957
15	O=C(O)CCC(=O)OC%24CC%16%26C8C=%10C=%12C%25=C%27C=%28C%13=C%29C=1C=%11C%14=C4C=1C%31=C%21C=3C=2C=%20C%17=C6C=5C=2C(C=34)=C%15C=9C=5C7=C6C=%18C(=C%16C7=C8C=9C(C=%10C=%11C=%12%13)=C%14%15)C=%23C%19C%22=C%32C(=C%17C=%18%19)C=%20C%21=C%30C%32=C(C%22=C%27C=%23C%25%26CC%24)C=%28C%29=C%30%31	304.4909654	5.200	6.363

(*Continued*)

Table 4. (*Continued*).

ID	SMILES	DCW(1)	pEC50 Expr	pEC50 Calc
18	O=C(O)CCCCCNC%31%30C%13=C1C%15=C%18C%26C1=C%28C%31C%10=C%29C=%20C9=C%21C2=C6C%22=C%24C=5C%17=C4C%16=C%14C8=C%12C=3C=%11C(=C2C=7C=3C8=C4C=5C6=7)C9=C%10C=%11C%30C%12=C%13C%14=C%15C%16=C%19C%17=C%23C(=C%18%19)C=%27C%25=C(C=%20C%21=C%22C%25=C%23%24)C(C%26=%27)=C%28%29	286.9281017	6.310	4.173
	Test set			
1	OC%22CC%21%24C%31=C5c1c%23c%12c%10c2c1C=4C3=C%32C2=C%11C=%25C=%33C=%26C%29=C8C(=C3C=7C(C=45)=C6C%31=C%20C=%18C9=C6C=7C8=C%30C9=C%19C%28=C%13C=%27C%14=C(c(c%10%11)c%15c%12C(=C%17C(=C(C%13C%17%16CC(O)CCC%14%15%16)C=%18%19)C%20%21)C%23%24CC%22)C=%25C=%26C=%27C%28=C%29%30)C%32=%33	292.5432013	4.750	4.873
9	O=C(O)C(N)CCCNC(=N)NC%31%30C%13=C1C%15=C%18C%26C1=C%28C%31C%10=C%29C=%20C9=C%21C2=C6C%22=C%24C=5C%17=C4C%16=C%14C8=C%12C=3C=%11C(=C2C=7C=3C8=C4C=5C6=7)C9=C%10C=%11C%30C%12=C%13C%14=C%15C%16=C%19C%17=C%23C(=C%18%19)C=%27C%25=C(C=%20C%21=C%22C%25=C%23%24)C(C%26=%27)=C%28%29	283.2251200	3.640	3.711
11	OCC(CO)(CO)NC(=O)CCCc1ccc(cc1)C%11(c2ccccc2)C%21%12C5=C%28C%20=C%19C%29=C%18C%31=C%17C=%32C=4C=3C=%16C%13=C7C=6C=3C=%22C=4C%34=C%23C=%25C=%33C%30=C(C%26=C5C=%24C%10C=9C%27=C(C=6C8=C7C=%14C(=C(C8=9)C%10%11%12)C(C%15C%19=C%18C(=C%13C=%14%15)C=%16%17)=C%20%21)C=%22C%23=C%27C=%24C=%25%26)C%28=C%29C%30=C%31C=%32C=%33%34	298.9331967	5.600	5.670
13	O=C(O)CO\N=C%26\CC%17%30C%14C=1C%27=C2C=%32C=1C%16=C%31C=5C=%32C=6C9=C2C%28=C%10C%24=C%23C8=C%22C=7C=4C=3C=%21C%18=C%12C=%11C=3C(C=4C=5C=6C=7C8=C9%10)=C%31C=%15C=%11C%13=C%12C=%19C(=C%17C%13=C%14C=%15%16)C=%25C%20C(=C%23C(=C%18C=%19%20)C=%21%22)C%24=C%29C=%25C%30(CC%26)C%27=C%28%29	300.0556465	5.530	5.810
17	O=C(O)CCC(=O)NCCc1ccc(cc1)C%11(c2ccc(CCNC(=O)CCC(=O)O)cc2)C%21%12C5=C%28C%20=C%19C%29=C%18C%31=C%17C=%32C=4C=3C=%16C%13=C7C=6C=3C=%22C=4C%34=C%23C=%25C=%33C%30=C(C%26=C5C=%24C%10C=9C%27=C(C=6C8=C7C=%14C(=C(C8=9)C%10%11%12)C(C%15C%19=C%18C(=C%13C=%14%15)C=%16%17)=C%20%21)C=%22C%23=C%27C=%24C=%25%26)C%28=C%29C%30=C%31C=%32C=%33%34	295.7213966	5.130	5.269

Split C, Threshold = 1, first probe of the Monte–Carlo optimization.

Journal of Computational Chemistry DOI 10.1002/jcc

Table 5. The Distribution of Attributes in the Subtraining Set, Calibration Set, and Test Set.

		Total number of S_k			Numbers of SMILES which are containing S_k		
S_k	No.	Subtraining set	Calibration set	Test set	Subtraining set	Calibration set	Test set
Split A							
%	1	328	370	242	7	8	5
(	2	94	118	82	7	8	5
/	3	0	1	1	0	1	1
0	4	42	48	30	7	8	5
1	5	196	224	140	7	8	5
2	6	192	218	140	7	8	5
3	7	96	104	78	7	8	5
4	8	44	50	36	7	8	5
5	9	44	48	30	7	8	5
6	10	42	48	30	7	8	5
7	11	42	48	30	7	8	5
8	12	42	48	30	7	8	5
9	13	42	48	30	7	8	5
=	14	252	302	198	7	8	5
@	15	0	2	0	0	1	0
C	16	455	547	332	7	8	5
H	17	0	2	0	0	1	0
N	18	9	8	4	4	6	3
O	19	24	22	12	6	8	4
[	20	0	4	0	0	1	0
\	21	2	5	1	1	3	1
c	22	32	22	36	3	2	3
Split B							
%	1	380	324	236	8	7	5
(	2	110	102	82	8	7	5
/	3	1	0	1	1	0	1
0	4	48	42	30	8	7	5
1	5	224	196	140	8	7	5
2	6	220	190	140	8	7	5
3	7	116	92	70	8	7	5
4	8	54	44	32	8	7	5
5	9	50	42	30	8	7	5
6	10	48	42	30	8	7	5
7	11	48	42	30	8	7	5
8	12	48	42	30	8	7	5
9	13	48	42	30	8	7	5
=	14	293	250	209	8	7	5
@	15	0	0	2	0	0	1
C	16	536	449	349	8	7	5
H	17	0	0	2	0	0	1
N	18	5	12	4	4	6	3
O	19	22	18	18	7	6	5
[	20	0	0	4	0	0	1
\	21	3	2	3	2	1	2
c	22	46	32	12	4	3	1
Split C							
%	1	378	324	238	8	7	5
(	2	102	102	90	8	7	5
/	3	2	0	0	2	0	0
0	4	48	42	30	8	7	5
1	5	224	196	140	8	7	5
2	6	220	192	138	8	7	5
3	7	114	90	74	8	7	5
4	8	52	44	34	8	7	5
5	9	50	42	30	8	7	5

(*Continued*)

Table 5. (*Continued*).

		Total number of S_k			Numbers of SMILES which are containing S_k		
S_k	No.	Subtraining set	Calibration set	Test set	Subtraining set	Calibration set	Test set
6	10	48	42	30	8	7	5
7	11	48	42	30	8	7	5
8	12	48	42	30	8	7	5
9	13	48	42	30	8	7	5
=	14	298	264	190	8	7	5
@	15	0	2	0	0	1	0
C	16	525	477	332	8	7	5
H	17	0	2	0	0	1	0
N	18	9	4	8	6	3	4
O	19	24	17	17	7	6	5
[	20	0	4	0	0	1	0
\	21	4	2	2	3	1	1
c	22	34	22	34	3	2	3

Conclusions

SMILES-based optimal descriptors can be predictor for the binding affinity values (pEC50) of fullerene derivatives; balance of correlations gives models that are more robust for the pEC50 than classic scheme (training set–test set); the difference in distribution of SMILES symbols in the subtraining set, calibration set, and the test set leads to the difference of statistical quality of models for three random splits examined in this study.

References

1. Durdagi, S.; Mavromoustakos, T.; Papadopoulos, M. G. Bioorg Med Chem Lett 2008, 18, 6283.
2. Durdagi, S.; Mavromoustakos, T.; Chronakis, N.; Papadopoulos, M. G. Bioorg Med Chem 2008, 16, 9957.
3. Castillo-Garit, J. A.; Martinez-Santiago, O.; Marrero-Ponce, Y.; Casan~ola-Marti'n, G. M.; Torrens, F. Chem Phys Lett 2008, 464, 107.
4. Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. QSAR Comb Sci 2006, 25, 928.
5. Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. Polymer 2006, 47, 3240.
6. Puzyn, T.; Mostrag, A.; Suzuki, N.; Falandysz, J. Atmos Environ 2008, 42, 6627.
7. Ray, S.; Sengupta, C.; Roy, K. Cent Eur J Chem 2008, 6, 267.
8. Roy, K.; Roy, P. P. Chem Biol Drug Des 2008, 72, 370.
9. Roy, K.; Ghosh, G. Chem Biol Drug Des 2008, 72, 383.
10. Toropov, A. A.; Toropova, A. P.; Gutman, I. Croat Chim Acta 2005, 78, 503.
11. Duchowicz, P. R.; Talevi, A.; Bruno-Blanch, L. E.; Castro, E. A. Bioorg Med Chem 2008, 16, 7944.
12. Duchowicz, P. R.; Vitale, M. G.; Castro, E. A. J Math Chem 2008, 44, 541.
13. Toropov, A. A.; Leszczynski, J. Chem Phys Lett 2006, 433, 125.
14. Toropov, A. A.; Leszczynska, D.; Leszczynski, J. Comput Biol Chem 2007, 31, 127.
15. Toropov, A. A.; Leszczynska, D.; Leszczynski, J. Chem Phys Lett 2007, 441, 119.
16. Toropov, A. A.; Leszczynska, D.; Leszczynski, J. Mater Lett 2007, 61, 4777.
17. Toropov, A. A.; Rasulev, B. F.; Leszczynska, D.; Leszczynski, J. Chem Phys Lett 2008, 457, 332.
18. Weininger, D. J Chem Inf Comput Sci 1988, 28, 31.
19. Weininger, D.; Weininger, A.; Weininger, J. L. J Chem Inf Comput Sci 1989, 29, 97.
20. Weininger, D. J Chem Inf Comput Sci 1990, 30, 237.
21. Vidal, D.; Thormann, M.; Pons, M. J Chem Inf Model 2005, 45, 386.
22. Toropov, A. A.; Toropova, A. P.; Mukhamedzhanova, D. V.; Gutman, I. Indian J Chem Sec A 2005, 44, 1545.
23. Doweyko, A. M. J Comput Aided Mol Des 2008, 22, 81.
24. Johnson, R. S. J Chem Inf Model 2008, 48, 25.
25. Golbraikh, A.; Tropsha, A. J Mol Graph Model 2002, 20, 269.
26. Roy, P. P.; Roy, K. QSAR Comb Sci 2008, 27, 302.
27. Toropov, A. A.; Rasulev, B. F.; Leszczynski, J. Bioorg Med Chem 2008, 16, 5999.
28. ACD/ChemSketch Freeware, version 11.00, Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2007. Available at: http://www.acdlabs.com.
29. U.S. Library of Medicine, 2008. Available at: http://toxnet.nlm.nih.gov/.
30. National Institute of Standard and Technology, 2008. Available at: http://webbook.nist.gov/chemistry/.